

DE CENTRALE LIMIETSTELLING

EEN BEWIJS VIA KARAKTERISTIEKE FUNCTIES

LETO CAVÉ

Abstract

De centrale limietstelling (CLS) verklaart waarom genormaliseerde sommen van onafhankelijke toevalsvariabelen asymptotisch normaal verdeeld zijn. Dit vormt de basis van de inferentiële statistiek. Dit artikel illustreert de CLS eerst met alledaagse voorbeelden, zoals de wachttijd bij een koffieautomaat, om de convergentie naar de klokvorm intuïtief te tonen. Vervolgens bewijzen we de stelling met karakteristieke functies en Taylorreeksen. We sluiten af met de implicaties voor betrouwbaarheidsintervallen en hypothesetoetsen.

Doelgroep: Bachelorstudenten wiskunde, statistiek, econometrie en natuurwetenschappen, en onderzoekers met voorkennis van mathematische statistiek en analyse.

SLEUTELWOORDEN: Centrale limietstelling, Wet van de grote getallen, Normale verdeling, Karakteristieke functie, Continuïteitsstelling van Lévy, Steekproefgemiddelde, Hypothesetoetsen

De centrale limietstelling

In de statistiek is de exacte kansverdeling van de populatie in de praktijk bijna altijd onbekend. Met behulp van een steekproef (Definitie 2) kunnen statistici een schatting maken van deze werkelijke verdeling.

De centrale limietstelling (CLS) laat zien dat het steekproefgemiddelde ($\bar{X}_n$) van een opeenvolging van onafhankelijke en identiek verdeelde (i.i.d.) stochasten met verwachting μ en eindige variantie $\sigma^2 < \infty$, asymptotisch normaal verdeeld is. Dit resultaat geldt onafhankelijk van de onderliggende verdeling van de stochasten:

$$\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1) \quad \Longleftrightarrow \quad \forall x \in \mathbb{R}, \quad \lim_{n \rightarrow \infty} P \left[\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq x \right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}t^2} dt = \Phi(x)$$

Waar de wet van de grote getallen (Stelling 7) garandeert dat het steekproefgemiddelde convergeert naar de populatieverwachting, beschrijft de centrale limietstelling de asymptotische verdeling.

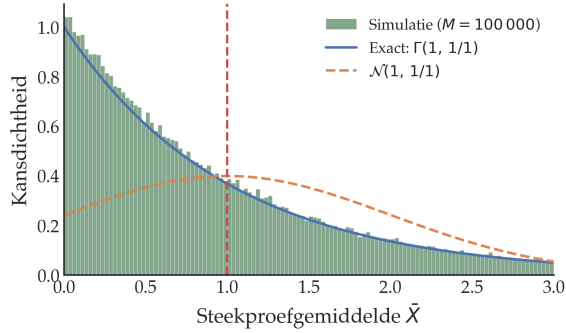
Wanneer we een steekproef $\{(X_i)\}_{i=1}^n$ van grootte n trekken uit een populatie (d.w.z. $X_i \stackrel{\text{i.i.d.}}{\sim} P$) en het steekproefgemiddelde berekenen, nadert de verdeling van dat gemiddelde de normale verdeling naarmate n toeneemt. Tegelijk daalt de variantie van het steekproefgemiddelde (σ^2/n), wat resulteert in een smallere dichtheidsfunctie en een nauwkeurigere schatting. Implicaties zoals het toetsen van hypothesen en het construeren van betrouwbaarheidsintervallen maken dat de centrale limietstelling onmisbaar is voor empirisch onderzoek.

Asymptotisch gedrag is direct te illustreren met een simulatie. Stel dat de wachttijd X bij een koffieautomaat exponentieel verdeeld is met parameter $\lambda = 1$ (dus $X \sim \text{Exp}(1)$), met $\mu = 1$ en $\sigma^2 = 1$. Een enkele steekproef van $n = 30$ observaties (gebaseerd op de vuistregel $n \geq 30$) levert een gemiddelde op van bijvoorbeeld 1.15. De theoretische populatieverdeling is sterk rechts-scheef. Voor deze verdeling is de exacte verdeling van het steekproefgemiddelde bekend: $\bar{X}_n \sim \Gamma(n, 1/n)$ (vorm n , schaal $1/n$), met scheefheid $2/\sqrt{n}$. Bij $n = 30$ bedraagt die scheefheid nog 0.365, wat laat zien dat de vuistregel $n \geq 30$ een vuistregel is en geen exact resultaat.

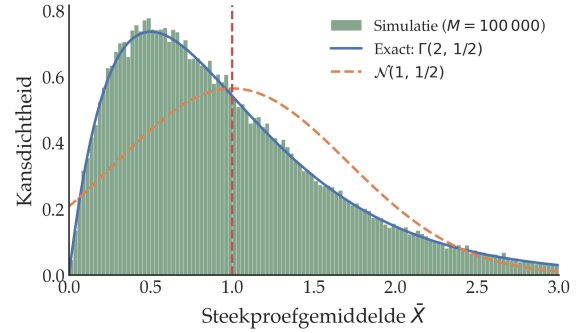
Het effect van de centrale limietstelling wordt zichtbaar wanneer we de steekproefgrootte n laten toenemen bij een vast aantal herhalingen M (zie Figuur 1).

Figuur 1

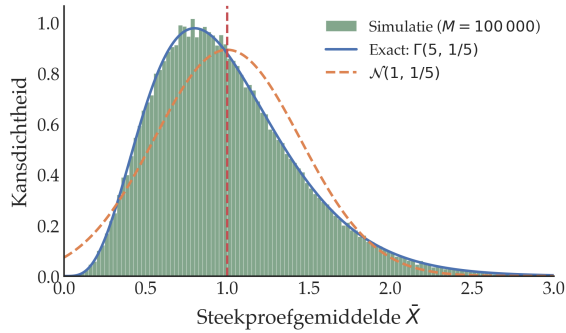
De verdeling van het steekproefgemiddelde bij toenemende steekproefgrootte n , telkens gebaseerd op $M = 100\,000$ herhalingen. De stippellijn is de normale verdeling $\mathcal{N}(1, 1/n)$, de doorgetrokken lijn de exacte verdeling $\Gamma(n, 1/n)$.



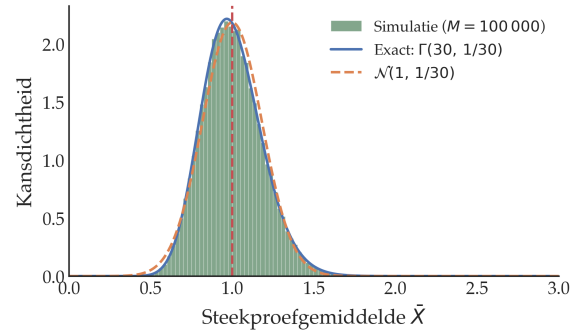
(a) $n = 1$: het steekproefgemiddelde is één waarneming; de verdeling is nog volledig scheef (scheefheid 2.00).



(b) $n = 2$: het gemiddelde van twee waarnemingen; de scheefheid is gehalveerd tot 1.41.



(c) $n = 5$: de klokvorm wordt zichtbaar (scheefheid 0.89).



(d) $n = 30$: de verdeling is vrijwel symmetrisch, maar een restscheefheid van 0.37 blijft.

In Figuur 1 valt op dat de steekproefgemiddelden zich met toenemende n steeds sterker concentreren rond de werkelijke verwachtingswaarde $\mu = 1$. Bij $n = 1$ is een waarde boven 3 nog goed mogelijk (kans $e^{-3} \approx 0.05$); bij $n = 30$ is een steekproefgemiddelde van 3 of 0.5 praktisch uitgesloten. Wanneer n toeneemt convergeert de verdeling van het steekproefgemiddelde naar de normale verdeling; M bepaalt slechts hoe scherp we die verdeling in beeld krijgen.

De stelling is niet beperkt tot continue verdelingen zoals de exponentiële verdeling. Beschouw bijvoorbeeld een discrete uniforme verdeling van een zuivere dobbelsteen, $X_i \sim \text{Unif}\{1, \dots, 6\}$, met $\mu = \frac{7}{2}$ en $\sigma^2 = \frac{35}{12}$. Voor de verdeling van de dobbelsteen is door symmetrie de normale benadering vaak al redelijk voor relatief kleine steekproeven.

Ook voor een Bernoulli-proces met $X_i \sim \text{Ber}(p)$ geldt de stelling. De som $S_n = \sum_{i=1}^n X_i$ is binomiaal verdeeld ($S_n \sim \text{Bin}(n, p)$), en de CLS resulteert hier in de klassieke stelling van De Moivre-Laplace (1733):

$$\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} \mathcal{N}(0, 1)$$

De aanname van een eindige variantie is hierbij echter strikt noodzakelijk. De Cauchy-verdeling bezit bijvoorbeeld geen gedefinieerde verwachtingswaarde of variantie. Als gevolg hiervan is de Lindeberg-Lévy-stelling niet van toepassing en convergeert het steekproefgemiddelde van Cauchy-stochasten niet naar een normale verdeling.

Formele definitie en het bewijs

Zij $(\Omega, \mathcal{F}, P)$ een kansruimte. Laat $X_1, X_2, \dots$ een rij i.i.d. stochastische variabelen zijn met

$$\mathbb{E}[X_i] = \mu \in \mathbb{R}, \quad \text{Var}[X_i] = \sigma^2 \in (0, \infty) \quad (\text{Eindige variantie})$$

Onder deze voorwaarden stelt de stelling van Lindeberg-Lévy [1] dat het gestandaardiseerde steekproefgemiddelde in de limiet standaardnormaal verdeeld is:

$$\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} = \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0, 1) \quad \text{als } n \rightarrow \infty \quad (\text{CLS})$$

Bewijs. We definiëren de nieuwe stochastische variabele $Y_i := \frac{1}{\sigma}(X_i - \mu)$, waarbij we opmerken dat

$$\mathbb{E}[Y_i] = 0, \quad \text{Var}[Y_i] = 1 \quad (\text{Standaardisatie})$$

Beschouw nu $Z_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i$. Omdat $Y_1, \dots, Y_n$ i.i.d. zijn, geldt voor de karakteristieke functie (Definitie 5) van Z_n :

$$\phi_{Z_n}(t) = \mathbb{E}[\exp(itZ_n)] = \mathbb{E}\left[\exp\left(it \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i\right)\right] = \prod_{i=1}^n \phi_{Y_i}\left(\frac{t}{\sqrt{n}}\right) = \left(\phi_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n$$

Via Taylorreeks (Stelling 6) schrijven we vervolgens ϕ_Y uit rond $s = 0$. Hierbij gebruiken we dat $\phi_Y^{(j)}(0) = i^j \mathbb{E}[Y^j]$.

$$\phi_Y(s) = 1 + is \underbrace{\mathbb{E}[Y]}_{=0} - \frac{s^2}{2} \underbrace{\mathbb{E}[Y^2]}_{=1} + o(s^2) = 1 - \frac{s^2}{2} + o(s^2), \quad s \rightarrow 0 \quad (\text{Taylorontwikkeling})$$

Merk op dat deze ontwikkeling tot tweede orde precies bestaat dankzij de aanname $\sigma^2 < \infty$; zonder eindige variantie bestaat $\phi_Y''(0)$ niet. Substitueer $s = t/\sqrt{n}$. Dan hebben we dat de karakteristieke functie van Z_n gelijk is aan $\phi_{Z_n}(t) = \left(1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n$. De standaardlimiet $\lim_{n \rightarrow \infty} \left(1 + \frac{a_n}{n}\right)^n = e^a$ voor $a_n \rightarrow a$ geldt ook voor complexe a_n , wat de stap rechtvaardigt ondanks dat ϕ_Y complexwaardig is. Nemen we nu de limiet $n \rightarrow \infty$, dan volgt:

$$\lim_{n \rightarrow \infty} \phi_{Z_n}(t) = e^{-t^2/2} = \phi_{N(0,1)}(t) \quad (\text{Karakteristieke functie van } N(0, 1))$$

Volgens de continuïteitsstelling van Lévy (Stelling 8)[1] is de puntsgewijze convergentie van karakteristieke functies naar een karakteristieke functie convergentie in distributie. Het volgt dat:

$$Z_n = \frac{1}{\sigma} \sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, 1) \quad \text{als } n \rightarrow \infty. \quad (\text{Convergentie met de Lévy continuïteitsstelling})$$

■

Toepassing in betrouwbaarheidsintervallen en hypothesetoetsen

De centrale limietstelling vormt de theoretische basis van vrijwel alle klassieke hypothesetoetsen (Definitie 11). Wanneer een steekproefgemiddelde voor grote n bij benadering normaal verdeeld is, kan men berekenen hoe waarschijnlijk een waargenomen afwijking van de nulhypothese is. Hierdoor kunnen onderzoekers p -waarden (Definitie 12) bepalen — dat is de kans dat we, onder de aanname van de nulhypothese, een volgende steekproefgemiddelde zouden vinden dat minstens zo extreem is als het waargenomen gemiddelde. Hoe kleiner deze kans is (vaak kiest men $p < 0.05$), hoe sterker ons bewijs om de nulhypothese te verwerpen ten gunste van de alternatieve hypothese.

Vrijwel iedere empirische wetenschap steunt impliciet op de centrale limietstelling. Wanneer een fysicus een meetonzekerheid rapporteert, een econoom een regressiecoëfficiënt toetst, een epidemioloog een betrouwbaarheidsinterval (Definitie 13) construeert of een farmaceut de werkzaamheid van een geneesmiddel evalueert, wordt er gebruikgemaakt van asymptotische normaliteit. De centrale limietstelling vormt daarmee een fundamentele overbrugging tussen eindige steekproeven en uitspraken over een onbekende populatie.

Het is echter van belang om op te merken dat deze normaliteit niet altijd voor lief kan worden genomen. Soms volgen observaties extreme machtswetten die wetenschappers vaak negeren door uitschieters te verwijderen of de data te transformeren, puur om de centrale limietstelling toch te kunnen toepassen. In bepaalde gevallen kan dit echter een grote fout zijn: zeer extreme, ogenschijnlijk onwaarschijnlijke gebeurtenissen zijn wel degelijk mogelijk en vormen in de maatschappij soms eerder de norm dan de uitzondering.

De centrale limietstelling is theoretisch zeer aantrekkelijk, maar in zulke situaties is het essentieel om alternatieve, robuustere methoden te overwegen. De impact van het hoogst onwaarschijnlijke, het onvoorspelbare en dat wat buiten onze verwachtingshorizon ligt, kan niet worden genegeerd. De natuur zit vol rampen die niemand zag aankomen en zelden wordt een financiële crash correct voorspeld. Zoals Nassim Taleb in zijn boek *De Zwarte Zwaan*[2] beschrijft, kunnen we in dit soort extreme gevallen niet vertrouwen op de centrale limietstelling — maar dat is een ontdekking voor een volgende keer.

Bronnen

- [1] Patrick Billingsley. *Probability and Measure*. 3rd ed. John Wiley and Sons, 1995.
- [2] Nassim Nicholas Taleb. *De Zwarte Zwaan: De impact van het hoogst onwaarschijnlijke*. Dutch. Uitgeverij Nieuwezijds, 2012. ISBN: 9789057123672.
- [3] Joseph C. Watkins. *An Introduction to the Science of Statistics: From Theory to Implementation*. Topic 11: The Central Limit Theorem. University of Arizona. URL: https://archive.math.arizona.edu/jwatkins/K_clt.pdf (visited on 09/04/2026).
- [4] Remco van der Hofstad. *Handout limietstellingen Kansrekening 2WS20*. Jan. 11, 2018. URL: https://rhofstad.win.tue.nl/hand_out_limit_theorems.pdf.
- [5] Jan van de Craats. *Kansrekening en Statistiek*. Dutch. Collegedictaat. Koninklijke Militaire Academie, Aug. 2002. URL: https://staff.fnwi.uva.nl/j.vandecraats/Kans_Stat.pdf.

Belangrijke Definities en Stellingen

Definitie 1 (Populatie)

Een populatie is de verzameling van eenheden waarover een uitspraak gedaan wordt. De verdeling van een kenmerk in de populatie wordt gekarakteriseerd door een (onbekende) parameter $\theta \in \Theta$ (de parameterruimte), zoals het gemiddelde μ en de variantie σ^2 , dus $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_{>0}$. In de praktijk is θ onbekend en wordt deze geschat op basis van een steekproef.

Definitie 2 (Steekproef $(X_1, \dots, X_n)$)

Een steekproef van omvang n is een rij stochasten $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} F$. Met i.i.d. bedoelen we dat de stochasten onafhankelijk en identiek verdeeld zijn volgens een gemeenschappelijke verdeling F . Onafhankelijkheid betekent dat voor alle $x_1, \dots, x_n \in \mathbb{R}$:

$$P(X_1 \leq x_1, \dots, X_n \leq x_n) = \prod_{i=1}^n F(x_i).$$

De gerealiseerde waarden $\{x_1, \dots, x_n\}$ heten de waarnemingen. Een functie $T(X_1, \dots, X_n)$ van de steekproef heet een steekproefstatistiek (bijv. $\bar{X}$ en S^2), en dient als schatter van de onbekende populatieparameter.

Definitie 3 (Steekproefgemiddelde $(\bar{X})$)

Het steekproefgemiddelde van $X_1, \dots, X_n$ is de stochast

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Als $\mathbb{E}[X_i] = \mu$ en $\text{Var}[X_i] = \sigma^2 < \infty$ voor alle i , dan $\mathbb{E}[\bar{X}] = \mu$ en $\text{Var}[\bar{X}] = \frac{\sigma^2}{n}$. Dit betekent ook dat $\bar{X}$ een zuivere schatter is van μ waarvan de variantie naar 0 gaat als $n \rightarrow \infty$.

Definitie 4 (Normale verdeling $(N(\mu, \sigma^2))$)

X volgt de normale (Gaussische) verdeling met parameters $\mu \in \mathbb{R}$, $\sigma^2 > 0$, genoteerd $X \sim N(\mu, \sigma^2)$, als X continu is met

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

De verdeling is symmetrisch rond μ ; σ^2 bepaalt de spreiding. Voor de standaardnormale verdeling ($\mu = 0, \sigma^2 = 1$), genoteerd $Z \sim N(0, 1)$, geldt de standaardisatie

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

Definitie 5 (Karakteristieke functie (ϕ_X))

De karakteristieke functie van X is $\phi_X : \mathbb{R} \rightarrow \mathbb{C}$ met

$$\phi_X(t) = \mathbb{E}[e^{itX}] = \mathbb{E}[\cos(tX)] + i \mathbb{E}[\sin(tX)], \quad t \in \mathbb{R}, \quad i = \sqrt{-1}.$$

Daarnaast geldt dat ϕ_X altijd bestaat ($|\phi_X(t)| \leq 1$) en de verdeling van X uniek bepaalt. Belangrijke eigenschappen zijn:

- voor $a, b \in \mathbb{R}$: $\phi_{aX+b}(t) = e^{itb} \phi_X(at)$;
- als $X \perp\!\!\!\perp Y$: $\phi_{X+Y}(t) = \phi_X(t) \phi_Y(t)$;
- voor $X \sim N(\mu, \sigma^2)$: $\phi_X(t) = \exp(i\mu t - \frac{1}{2}\sigma^2 t^2)$.

Stelling 6 (Stelling van Taylor (met kleine-o restterm))

Zij f n keer differentieerbaar in een punt a . Dan geldt voor $x \rightarrow a$:

$$f(x) = \sum_{k=0}^n \frac{f^{(k)}(a)}{k!} (x-a)^k + o(|x-a|^n).$$

Zoals we in het artikel zien passen we dit toe rond $a = 0$:

$$f(t) = \sum_{k=0}^n \frac{f^{(k)}(0)}{k!} t^k + o(|t|^n).$$

Dit betekent dat de restterm sneller naar nul convergeert dan $|t|^n$, oftewel $\lim_{t \rightarrow 0} \frac{R_n(t)}{|t|^n} = 0$.

Stelling 7 (Wet van de Grote Getallen (WGG))

Zij $X_1, X_2, \dots$ i.i.d. met $\mathbb{E}[|X_i|] < \infty$ en $\mathbb{E}[X_i] = \mu$, en $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

Zwakke WGG. Voor elke $\varepsilon > 0$:

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0, \quad \text{we noteren } \bar{X}_n \xrightarrow{P} \mu \quad (\text{convergentie in kans})$$

Sterke WGG.

$$P\left(\lim_{n \rightarrow \infty} \bar{X}_n = \mu\right) = 1, \quad \text{we noteren } \bar{X}_n \xrightarrow{a.s.} \mu \quad (\text{bijna zekere convergentie})$$

Belangrijk om te weten is dat de sterke WGG de zwakke impliceert. Intuïtief is het dat $\bar{X}_n$ convergeert naar μ als $n \rightarrow \infty$.

Stelling 8 (Continuïteitsstelling van Lévy)

Zij $X_1, X_2, \dots$ stochasten met karakteristieke functies ϕ_{X_n} , en X een stochast met karakteristieke functie

ϕ_X . Als $\phi_{X_n}(t) \rightarrow \phi_X(t)$ voor alle $t \in \mathbb{R}$ wanneer $n \rightarrow \infty$, dan geldt

$$X_n \xrightarrow{d} X \quad (\text{convergentie in verdeling})$$

Dit helpt om de CLS te bewijzen: men toont aan dat de karakteristieke functie van $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ puntsgewijs convergeert naar $\phi(t) = e^{-t^2/2}$, de karakteristieke functie van $N(0, 1)$.

Stelling 9 (Centrale limietstelling (CLS))

Zij $X_1, \dots, X_n$ i.i.d. met $\mathbb{E}[X_i] = \mu$ en $\text{Var}[X_i] = \sigma^2 \in (0, \infty)$. Dan

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0, 1) \quad (n \rightarrow \infty),$$

Definitie 10 (Hypotheses (H_0, H_1))

Zij $\theta \in \Theta$ de onbekende parameter. Een toets verdeelt Θ in disjuncte ruimtes Θ_0 en Θ_1 met $\Theta = \Theta_0 \cup \Theta_1$:

$$\begin{aligned} H_0 : \theta \in \Theta_0 & \quad (\text{nulhypothese}), \\ H_1 : \theta \in \Theta_1 & \quad (\text{alternatieve hypothese}). \end{aligned}$$

H_0 representeert de conservatieve aanname; H_1 is de bewering die op basis van de data al dan niet ondersteund wordt. Bestaat Θ_0 (of Θ_1) uit precies één waarde, dan spreekt men van een enkelvoudige hypothese, anders noemen we deze samengesteld.

Definitie 11 (Hypothesetoetsen)

Een hypothesetoets is een beslissingsregel gebaseerd op een toetsingsgrootheid $T = T(X_1, \dots, X_n)$. We verwerpen H_0 als T in het kritieke gebied $C \subseteq \mathbb{R}$ ligt:

$$\text{Verwerp } H_0 \iff T \in C.$$

Het significantieniveau α is de kans op een type I-fout (we verwerpen dan ten onrechte H_0):

$$\alpha = P(T \in C \mid H_0 \text{ waar}).$$

C wordt zo gekozen dat deze kans gelijk is aan een gekozen α (bijv. 0.05). Een type II-fout treedt op als H_0 niet verworpen wordt terwijl H_1 waar is.

Definitie 12 (p -waarde)

Zij t_{obs} de geobserveerde waarde van T . De p -waarde is de kans, berekend onder H_0 , om een waarde van T te observeren die minstens zo extreem is als t_{obs} . Voor een rechtszijdige toets geldt:

$$p = P(T \geq t_{\text{obs}} \mid H_0).$$

Andersom hebben we voor een linkszijdige toets $p = P(T \leq t_{\text{obs}} \mid H_0)$ en voor een tweezijdige toets $p = 2 \min\{P(T \geq t_{\text{obs}} \mid H_0), P(T \leq t_{\text{obs}} \mid H_0)\}$. Ook kan men hiermee een beslissingsregel vaststellen: verwerp H_0 als $p < \alpha$ — equivalent aan $t_{\text{obs}} \in C$. Let wel op want een p -waarde is geen kans dat H_0 waar is; het is een uitspraak over de data gegeven H_0 .

Definitie 13 (Betrouwbaarheidsinterval)

Een $(1 - \alpha) \times 100\%$ betrouwbaarheidsinterval voor θ is een stochastisch interval $[L(X_1, \dots, X_n), U(X_1, \dots, X_n)]$ zodanig dat, voor de vaste waarde θ :

$$P(L \leq \theta \leq U) = 1 - \alpha.$$

Deze gelijkheid geldt per definitie exact, terwijl het onderstaande CLS-interval die dekking pas asymptotisch bereikt. Voor μ met bekende σ^2 geeft de CLS bij benadering het interval $\left[\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right]$ met $z_{\alpha/2}$ het $(1 - \alpha/2)$ kwantiel van $N(0, 1)$. Dit betekent dat wanneer L en U stochastisch zijn, θ vast staat. Wanneer we spreken over een betrouwbaarheid van 95% zeggen we dat 95% van de gegenereerde intervallen de ware parameter zal bevatten. Vergis je niet: het is fout om te veronderstellen dat dit de kans is dat θ in één gerealiseerd interval ligt.